On Learning Frequency-Instance Correlations by Model-Agnostic Training for Synthetic Speech Detection

Zining Wang 2222208041@STMAIL.UJS.EDU.CN
School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
Lijian Gao LJGAO@STMAIL.UJS.EDU.CN
School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
Jialin Zhang 2212108042@STMAIL.UJS.EDU.CN
School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
Qirong Mao* MAO_QR@UJS.EDU.CN
School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
Jiangsu Engineering Research Center of Big Data Ubiquitous Perception and Intelligent Agriculture Applications, Zhenjiang, China
Provincial Key Laboratory of Computational Intelligence and New Technologies in Low-Altitude Digital Agriculture, Zhenjiang, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

The goal of Synthetic Speech Detection (SSD) is to detect spoofing speech synthesized by text-to-speech and voice conversion. Most existing SSD methods focus only on mining frequency-wise dependency by customizing frequency-aggregation modules in SSD models. However, the instance-wise dependency is usually under-explored, which is critical for identifying the synthetic speech from a global view. In this paper, we propose a novel model-agnostic training strategy for SSD that exploits both local (frequency-wise) and global (instance-wise) contexts, which do not rely on a customized architecture and can be flexibly integrated into previous SSD models. Specifically, we propose an inter-frequency correlation module to capture the local context by reconstructing the masked frequency information from the unmasked frequency context. Meanwhile, an inter-instance correlation module is performed to explore the global context among different instances by promoting intra-class compactness and inter-class dispersion in the latent space. These two complementary modules operate from distinct contextual perspectives, leading to improvements in SSD performance. Extensive experiments show that our method significantly improves the performance of two state-of-the-art models on the 2019 dataset and 2021 dataset of ASVspoof.

Keywords: Anti-spoofing; synthetic speech detection; consistency loss

1. Introduction

Synthetic Speech Detection (SSD) is the process of determining whether a given utterance is synthetic speech. With the advancement of deep learning technologies such as AI-Generated Content (AIGC) (Wen et al., 2023), Text-To-Speech (TTS) (Kim et al., 2022) and Voice Conversion (VC) (Li et al., 2023), artificially generated speeches pose a threat to automatic speaker verification systems(Qin et al., 2023; Yao et al., 2023). SSD plays a critical role

* Corresponding author

in protecting user privacy, preventing telecom fraud, enhancing the reliability of speech systems, and improving the user experience.

A mainstream of speech processing delves into network designs for time-frequency representation learning, which aims to capture local context (Gao et al., 2024, 2023). As shown in Fig. 1(a), existing methods often customize the network architecture for effective feature aggregation. For example, Convolutional Neural Network (CNN) is utilized to capture the discriminative time-frequency features (Tak et al., 2020). To capture the relationships between different sub-bands (e.g. spoofing artifacts present simultaneously in two different sub-bands), Tak et al. (2021) model the non-Euclidean data manifold spanning different sub-bands and temporal segments by Graph Neural Networks (GNN). Subsequent effort incorporates various techniques such as Transformer (Liu et al., 2023), to capture long-range contextual information. On the other hand, the significant improvement of auxiliary features (Kim and Ban, 2023) or performing model fusion (Zhang et al., 2023) demonstrates their effectiveness in the SSD task. To learn more discriminative time-frequency features, these approaches usually rely on architecture customization to model local context. A question is naturally raised: **Can we model local context in a model-agnostic manner?**

Recently, exploiting the relationship among different instances has achieved remarkable success in video representation learning tasks. Park et al. (2022) treat a segment of a video as an instance, considering different video segments as different perspectives. By pulling same instances closer together and pushing different instances farther apart in the feature space, the model is encouraged to learn the spatio-temporal features of the video through a self-supervised manner. This approach takes into account the rich semantic relationships between different video frames from a global view, motivating us to explore the role of relationships between speech samples in SSD. We raise another question: **For SSD task, can we model global context by exploiting the instance-wise relationship?**

To answer the above questions, we propose a novel model-agnostic training approach for SSD that exploits both local and global contexts. Specifically, we introduce an Inter-Frequency (InF) correlation learning module based on band-pass filtering and consistency loss. This module enhances the model's ability to learn discriminative features across different frequency bands. Moreover, we introduce an Inter-Instance (InI) correlation learning module that aims to bring instances of the same class closer together and push instances from different classes farther apart in the feature space, thus achieving intra-class consistency and inter-class disparity. The InF and InI modules can be seamlessly incorporated into existing SSD networks, which leads to performance gains without extra computation cost during inference.

In summary, our contributions can be summarized as follows:

- We propose a novel model-agnostic training strategy exploiting both local and global contexts (i.e., frequency-instance correlations), which can be easily plugged into existing SSD models.
- We propose the InF module to capture discriminative features from different frequency bands and the InI module to encourage intra-class consistency and inter-class differences.
- On the basis of LCNN (Das, 2021) and AASIST (Jung et al., 2022) models, our training method achieves SOTA performance both on the 2019 dataset (Wang et al.,

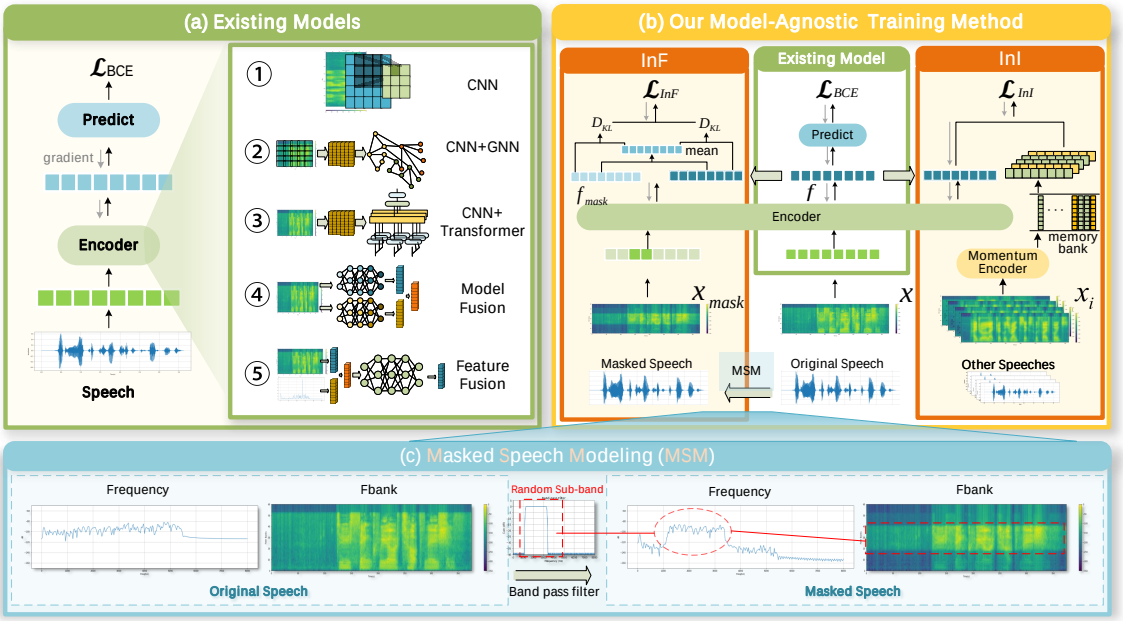

Figure 1: **Illustration of existing SSD models and our model-agnostic training method.** **(a)** Existing methods mainly focus on architecture customization, including ① CNN (Tak et al., 2020), ② CNN+GNN (Jung et al., 2022), ③ CNN+Transformer (Liu et al., 2023), ④ model fusion (Zhang et al., 2023), and ⑤ feature fusion (Kim and Ban, 2023). **(b)** By designing the InF and InI modules, our model-agnostic method can be flexibly plugged into previous SSD architectures for performance gains. The InF module is utilized to learn the correlation between frequency, while the InI module is employed to learn the correlation between instances. By randomly masking original speech via Masked Speech Modeling (MSM), our InF module would implicitly learn to model the masked frequency information under the guidance of a consistency loss. **(c)** By employing band-pass filters to filter the speech signals, MSM preserves signals within a designated frequency range while suppressing signals outside of that range.

2020) and 2021 dataset (Yamagishi et al., 2021) of ASVspoof. We further validate the effectiveness of each module through experimental results and visualizations.

2. Related work

In this section, we first give a brief overview of the system architecture for the SSD task. After that, a description and an analysis of the inter-frequency correlation and inter-instance correlation in the SSD task will be presented.

2.1. Systems for synthetic speech detection

Using deep neural networks for synthetic speech detection is broadly divided into three parts: (1) computing manual features which rely on theoretical and empirical understanding of speech signals, (2) extracting embeddings using neural networks, which can learn more complex feature representations and capture high-level semantic information of speech, (3) classifiers and loss function.

The manual feature extraction approach leverages human auditory perception characteristics to effectively extract critical information from speech. This extracted information effectively represents the distinctive features in the speech signal, ensuring robustness even in noisy environments. Inspired by the speaker verification task, synthetic speech detection also tries to use manual features as model inputs, such as MFCC, CQCC, FBank, and so on.

To extract deeper features, the manual features are fed into the model. By leveraging the learning of multi-layer networks, we can extract more critical features from high-dimensional data. This automatic feature learning enables efficient results even with limited data resources.

Finally, the embedding features extracted by the model are fed into the classifier and loss function for constraints. Commonly used loss functions are BCE loss, OC-Sotfmax loss, and MSE loss.

2.2. The inter-frequency correlation in synthetic speech detection

There are differences in the artifactual information carried by different frequency bands of synthetic speech. For example, on the ASVspoof2019 dataset, it turns out that the high-frequency features cause the system to overfit, while the low-frequency features are more robust but less accurate against known attacks (Zhang et al., 2021). However, on the 2021 data set, it is found that the difference between the high frequency components of the spoofed speech and the real speech remained discriminatory even after transmission and encoding in VOIP and PSTN systems (Huang et al., 2023). Therefore, it is a worthwhile research problem to help models learn more general and comprehensive band information.

Moreover, there are many improved methods like data augmentation that have been studied. For example, based on prior knowledge of test data and specific telephony scenarios, Rawboost (Tak et al., 2022) improves the detection performance of the model by adding linear and nonlinear noise to the training data. DASC (Das et al., 2021) uses signal companding techniques based on a-law and mu-law. In this method, the signal is first compressed and then expanded, which is widely used in telephone, speech, and many other audio applications. The above methods increase the diversity of sample band information by data augmentation. However, these are designed for the telephone scenarios or noise environment of the test data. Therefore, in non-telephony scenarios, the enhancement effect may be sub-optimal. Furthermore, a robust model should produce reliable and accurate outputs even when encountering minor perturbations in the input data. In the domain of Natural Language Processing (NLP), it has been demonstrated that constraining minor perturbations caused by data augmentation is beneficial to the model (Qiang et al., 2024). However, the existing enhancement methods for SSD do not consider the constraint on such disturbances.

Therefore, we propose sub-band augmentation and consistency loss for synthetic speech detection. It can encourage the model to consider a broader range of features to reduce overfitting, which over-focuses on a few key features in the training data that may not always be present, causing the model to perform poorly on new data. Consistency loss can constrain the model to reduce its sensitivity to data shift and enforce smoother neural network responses.

2.3. The inter-instance correlation in synthetic speech detection

In addition to learning rich time-frequency information, exploiting the relationships between multiple instances can also improve the effectiveness of the model. In the domain of videos, Park et al. (2022) treat a segment of a video as an instance, considering different video segments as different perspectives. By pulling same instance closer together and pushing different instances farther apart in the feature space, the model is encouraged to learn the spatio-temporal features of the video through a self-supervised manner. This approach takes into account the rich semantic relationships between different video frames. However, in the domain of SSD, limited consideration is given to the correlations between instances.

Therefore, we propose an inter-instance learning module. To compare samples, we set the memory bank to store historical samples. This module reduces the intra-class distance and increases the inter-class distance.

3. Proposed Method

Fig. 1(b) demonstrates our model-agnostic SSD approach, which integrates InF and InI modules. In the InF module, the input x is randomly masked using a Masked Speech Modeling (MSM) strategy based on band-pass filtering. The original speech and the masked speech are regularized to be consistent in the embedding space. In the InI module, speech samples from each batch undergo feature extraction by the momentum encoder and are subsequently stored in the memory bank. The embedded samples in each training batch are contrastively drawn closer to those historical samples with similar semantics.

3.1. Learning inter-frequency correlation

Previous studies have revealed that the artifacts introduced by various spoofing algorithms commonly manifest in distinct frequency ranges (Nautsch et al., 2021), so the SSD model may rely heavily on specific sub-bands to predict synthetic speech. To capture relevant artifact information, one may explicitly leverage the frequency-wise relationships within the spectrogram of a speech signal (Chen et al., 2023), thereby benefitting feature learning for accurate synthetic speech detection.

3.1.1. Masked Speech Modeling Based on Band-pass Filtering

Recently, Masked Image Modeling (MIM) has excelled at learning visual representations by exploiting the correlation among neighboring pixels (Chen et al., 2024; He et al., 2022). Inspired by this, several efforts in SSD have directly adopted the MIM concept, dividing the spectrogram of a speech signal into structural patches to facilitate the learning of time-frequency correlation. However, the manually extracted magnitude spectrogram may

potentially result in a loss of valuable information (e.g. phase information) (Tak et al., 2020). Another alternative involves the use of low-pass and high-pass filters (Tomilov et al., 2021) to directly mask the raw waveform, which can emulate the interference caused by codecs. Nevertheless, due to the characteristics of low-pass and high-pass filters, the masked frequency range is often restricted.

To effectively model the frequency-wise relationships (local context), we propose a novel MSM method based on band-pass filtering (in Fig. 1(c)), which selectively activates frequency features within a specific sub-band. MSM can significantly improve the randomness of masking by randomly setting a 2kHz frequency band within the 0-8kHz range. Technically, we use the Remez algorithm (Karam and McClellan, 1994) to construct a finite impulse response band-pass filter. The sampling frequency is 16000Hz. The transition bandwidth of the band-pass filter is set to 160Hz. The minimum frequency is randomly selected between 160Hz and 5840Hz. The maximum frequency is set to 2000Hz above the minimum frequency. The described filter can allow frequency features of random sub-band with 2000Hz width to pass through, while suppressing other frequency band features. The specific way of filtering with the original waveform is as follows: $x_{mask} = x * h$, where x_{mask} represents the signal after discrete convolution, x represents the original signal, and h represents the filter.

3.1.2. Kullback-Leibler (KL) Divergence Consistency Loss

A naive way is to leverage the masked speech x_{mask} as the augmented sample for model training, employing the cross-entropy loss. But it may fail to exploit the connection between the masked speech x_{mask} and the original speech x. To tackle this issue, we encourage the model to generate consistent features for the paired speech input by implementing the Kullback-Leibler (KL) divergence consistency loss. In that way, the model is devoted to recovering the masked frequency information by solely utilizing the unmasked frequency context, thus modeling the frequency-wise relationship.

The specific formula for KL Divergence is as follows:

$$D_{KL}(P||Q) = \sum_{i} P(i) log(\frac{P(i)}{Q(i)}) \tag{1}$$

The formula $D_{KL}(P||Q)$ represents how much information is lost when probability distribution Q is fitted to probability distribution P.

And the final consistency loss is as follows:

$$\mathcal{L}_{InF} = -\frac{1}{N}\sum_{i=1}^{N}[\frac{1}{2}(D_{KL}(q_i||\frac{q_i + q_i'}{2}) + D_{KL}(q_i'||\frac{q_i + q_i'}{2}))] \tag{2}$$

Here, q_i and q_i' are probability distributions calculated by the classifier from the embedding features f_i and its masked version f_{masked}, respectively. N denotes the number of samples. This consistency loss encourages the model to maintain stability and insensitivity to inputs across different frequency ranges.

3.2. Learning inter-instance correlation

In addition to focusing on frequency-wise correlation (local context), it is also essential to consider instance-wise correlation (global context). For example, different bonafide speeches often exhibit similarities in terms of silent segments (Zhang et al., 2021). To this end, different embedded bonafide speeches should be clustered together in the embedding space. To achieve this, we propose an inter-instance correlation module that pulls samples in the same category closer and pushes different samples in different categories farther apart in the feature space.

A straightforward approach is to learn inter-instance correlation in batch. Due to the limited number of instances in batch and to avoid excessive memory consumption, we employ a memory bank to store historical instances. The features stored in the memory bank come from the momentum encoder (He et al., 2020), and its weights update follows the momentum update rule as shown below.

$$\theta_k \leftarrow a\theta_k + (1-a)\theta_q \tag{3}$$

Here, θ_k denotes the parameters of the momentum encoder, and θ_q denotes the parameters of the encoder. $a \in [0, 1)$ is a momentum coefficient.

We use labels to construct positive and negative instance pairs. The $f^+ \in F^+$ denotes the positive sample features in the memory bank of the f_i. And the $f^- \in F^-$ denotes the negative sample features. Therefore, the similarity loss function is as follows.

$$\begin{aligned}\mathcal{L}_{InI} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\frac{1}{N_{pos}}\sum_{f^+}\hat{y}log\Big(\frac{1}{1+e^{(-cosine(f_i,f^+))}}\Big) \\ + \frac{1}{N_{neg}}\sum_{f^-}(1-\hat{y})log\Big(1-\frac{1}{1+e^{(-cosine(f_i,f^-))}}\Big)\Big]\end{aligned} \tag{4}$$

Here, *cosine* denotes the cosine similarity loss. N_{pos} and N_{neg} denote the number of positive and negative sample features, respectively. For the positive pairs, the $\hat{y} = 1$. And $\hat{y} = 0$ for the negative pairs.

The InF module focuses on the frequency relationships within the instances, while the InI module focuses on the relationships between instances. These modules are complementary and compatible. The classification loss we used is the Binary Cross-Entropy loss (BCE loss), which is one of the most commonly used loss in the field.

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(g(f_i)) + (1-y_i)log(1-g(f_i))] \tag{5}$$

Here, $f_i \in \mathcal{R}^N$ is the embedding extracted by model. And y_i is the label of instance. The final classification loss is as follows.

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{InF} + \lambda_2\mathcal{L}_{InI} + \mathcal{L}_{BCE}(x, y) + \mathcal{L}_{BCE}(x_{mask}, y) \tag{6}$$

where λ_1 and λ_2 are scalar used to balance $\mathcal{L}_{InF}$, $\mathcal{L}_{InI}$ and classification loss.

4. Experiments

4.1. Datasets and Metrics

All experiments are conducted using the ASVspoof2019 (Wang et al., 2020) and ASVspoof-2021 (Yamagishi et al., 2021) datasets. The model performance is evaluated using the Equal Error Rate (EER) and the Minimum Tandem Detection Cost Function (min t-DCF) (Kinnunen et al., 2020).

4.2. Feature Extraction and Implementation Details

For the LCNN model, the input speech length is fixed at 6 seconds. Then, we use Fbank with the first-order dynamic feature and second-order dynamic feature as the input during all the experiments. The input speech is sampled at a sampling rate of 16000Hz. The frame length is 512 sampling points and the frame shift is 128 sampling points. For the AASIST, 6 seconds of raw waveform data is used as input.

We implement our model in PyTorch, using the Adam (Kingma and Ba, 2015) optimizer, setting the β_1 parameter to 0.9 and the β_2 parameter to 0.999 to update the weights in the model. For the LCNN, the batch size is set to 64, and for the AASIST, the batch size is set to 8. The learning rate is initially set to 0.0003 and decays by 50% every 10 epochs. For $\mathcal{L}_{InF}$ and $\mathcal{L}_{InI}$, $\lambda_1 = 0.1$, $\lambda_2 = 1$ during LCNN+InF+InI training, and $\lambda_1 = 0.1$, $\lambda_2 = 0.0001$ during AASIST+InF+InI training. We train 100 epochs on 4 NVIDIA GTX 3080 Ti GPUs. The model with the lowest validation EER is then selected for evaluation.

4.3. Ablation Experiments

This section mainly proves the universality and effectiveness of our method by analyzing the experimental results. The main ablation experimental results of our model-agnostic are shown in Table 1. We did ablation experiments on two datasets of the ASVspoof using the LCNN model based on hand-crafted features and the end-to-end model AASIST. The t-SNE visualizations of features learned from method LCNN and our method are shown in Fig. 2.

4.3.1. Learning from inter-frequency correlation

InF denotes the incorporation of frequency band filter masking and consistency loss during the model training process. According to the experimental results in Table 1, LCNN+InF reduces EER by 41.9% (4.22 →2.45) on the 2019 test set and 30.1% (5.77 →4.03) on the 2021 test set compared to LCNN. Similarly, AASIST+InF also showed reductions in EER and min t-DCF compared with AASIST. The role of InF during the training process is evident not only in the improvement of experimental results but also in the spatial distribution of features. Observing the distribution of different frequency features of the same speech (in Fig. 2(a) red points), it is evident that the baseline model fails to represent them consistently. In contrast, our proposed training method effectively achieves this consistency in representation (in Fig. 2(b) red points).

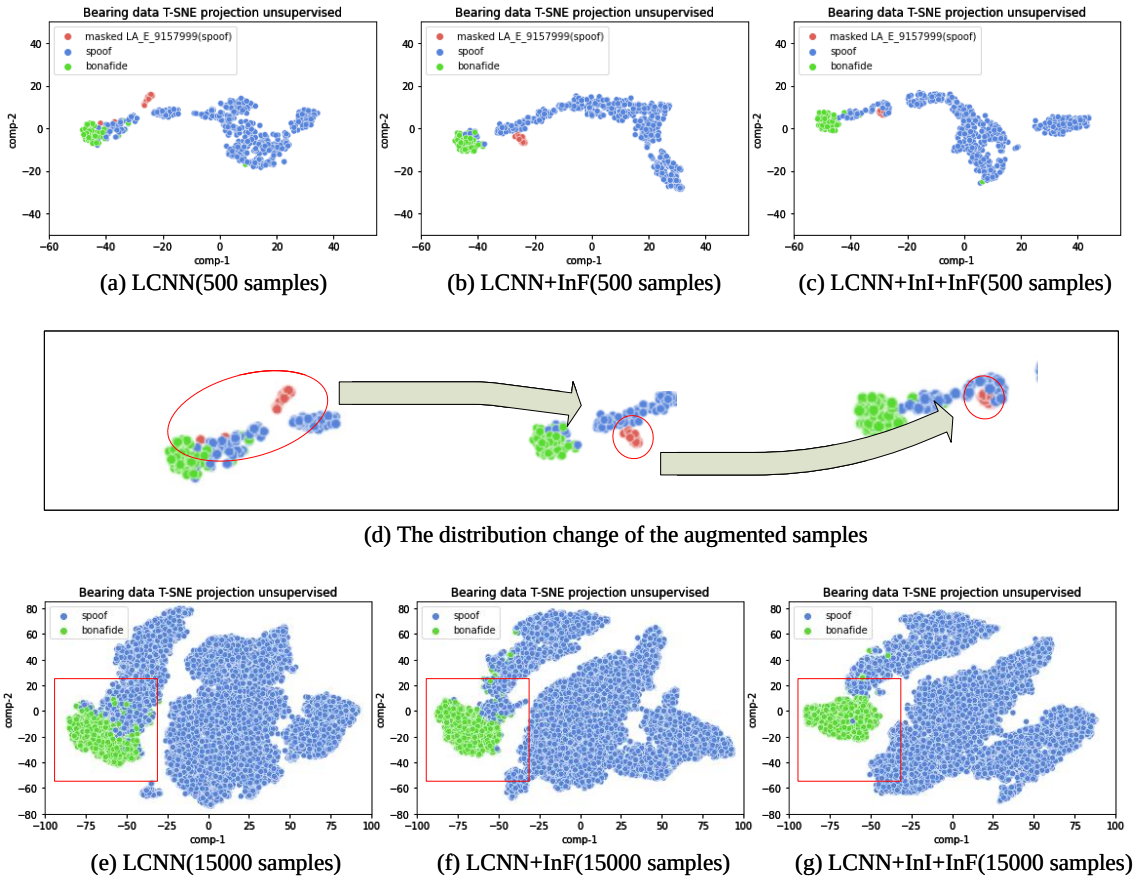

Figure 2: The t-SNE (Van der Maaten and Hinton, 2008) visualizations of some features learned from method LCNN and our methods. Green points represent genuine speech, blue points represent spoofed speech, and red points correspond to 20 samples generated from spoofed sample LA_E_9157999 using random filter masking. (a)(b)(c) represent the distribution of 500 samples and 20 augmented samples on the LCNN model and the improved methods. (d) shows the effects of the InF and InI modules on 20 augmented samples in detail. As can be seen in the red circle, the LCNN is unable to produce similar representations for different masked samples. Adding the InF module yields similar embeddings; After adding the InI module, the distance between the enhanced features and other spoofed samples is narrowed. (e)(f)(g) represent the distribution of 15000 samples on the 2019 test set.

4.3.2. Learning from inter-instance correlation

InI denotes the incorporation of cosine similarity loss during the training process. The experiments in Table 1 show that the InI module reduces the EER of the models on multiple datasets. LCNN+InI reduces EER by 53.5% (4.22 →1.96) on the 2019 test set and 35.7%

Table 1: Ablation Experiments. Each set of experiments is repeated three times with seed values of 1, 100, and 1000, and the final results are then averaged.

System	**ASVspoof2019**		**ASVspoof2021**	
	EER(%)	t-DCF	EER(%)	t-DCF
LCNN (Das, 2021)	4.22	0.1073	5.77	0.3107
LCNN+InF	2.45	0.0675	4.03	0.2813
LCNN+InI	1.96	0.0592	3.71	0.2720
LCNN+InF+InI	1.90	0.0550	3.15	0.2611
AASIST (Jung et al., 2022)	1.04	0.0317	6.24	0.3428
AASIST+InF	0.64	0.0205	5.19	0.3138
AASIST+InI	0.95	0.0315	5.08	0.3127
AASIST+InF+InI	0.62	0.0205	4.65	0.3017

Table 2: Performance comparison with existing single systems on the evaluation set of the ASVspoof 2019/2021 LA scenario.

System	**ASVspoof2019**		**ASVspoof2021**	
	EER(%)	t-DCF	EER(%)	t-DCF
DFSincNet (Huang et al., 2023) [SPL23]	0.52	0.0176	3.05	**0.2601**
GST+GCN (Chen et al., 2023) [ICASSP23]	0.58	0.0166	-	-
Rawformer (Liu et al., 2023) [ICASSP23]	0.59	0.0184	4.53	0.3088
ECANet (Xue et al., 2023) [ICASSP23]	0.88	0.0295	-	-
Res2Net (Kim and Ban, 2023) [2023]	0.94	0.0270	-	-
OCT (Li et al., 2022) [SPL22]	1.06	0.0345	-	-
AASIST (Jung et al., 2022) [ICASSP22]	0.83	0.0275	5.82	0.3349
AASIST+InF+InI(Ours)	**0.49**	**0.0160**	3.51	0.2728
LCNN (Das, 2021) [ASVspoof21]	3.37	0.0875	4.94	0.2979
LCNN+InF+InI(Ours)	1.80	0.0523	**3.01**	0.2602

(5.77 →3.71) on the 2021 test set compared to LCNN. Moreover, from the perspective of spatial feature distribution, our method exhibits better classification performance with

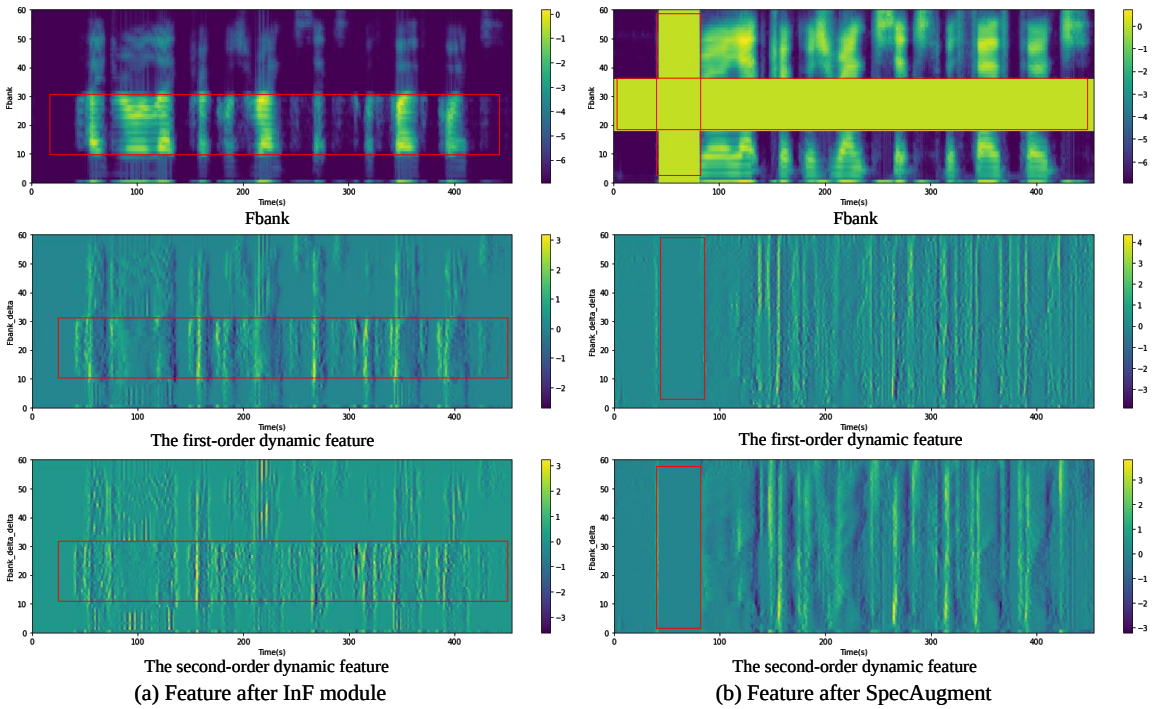

Figure 3: Differences between SpecAugment and InF methods. (a) is the feature map after InF filtering masking. Inside the red box is the frequency band passed after filtering. And (b) is the feature map after SpecAugment (Park et al., 2019). Inside the red box is the masked frequency band.

more clear boundaries (in Fig. 2(e)(f)(g)). It is demonstrated that training with the InI module can enhance the model's ability to distinguish more challenging instances. At the same time, by comparing Fig. 2(b) with Fig. 2(c), it can be seen that after adding the InF module, the model recognizes several spoofed speeches (blue points) from bonafide speech (green points) distribution.

4.4. Comparison with Other Systems

Table 2 presents the best results achieved by our single system as well as the results from other methods. In contrast to architecture customization models, our method has higher generality and works with most of the models. AASIST+Inf+InI achieves an EER of 0.49% on the 2019 test set, which is the best result among the current models. LCNN+InF+InI also achieves a competitive result with an EER of 3.01% on the 2021 test set. While our model-agnostic approach improves the detection of the model steadily, it is unable to achieve optimality on multiple datasets at the same time, as DFSincNet (Huang et al., 2023) does. In our perspective, this issue is correlated with the selection of baseline models and data augmentation techniques. The performance of the AASIST model itself appears to be

Table 3: Performance comparison with SpecAugment. Each set of experiments is repeated three times with seed values of 1, 100, and 1000, and the final results are then averaged.

System	seed	**ASVspoof2019**		**ASVspoof2021**	
		EER(%)	t-DCF	EER(%)	t-DCF
LCNN (Das, 2021)	1	3.60	0.0908	5.25	0.3023
	100	3.37	0.0875	4.94	0.2979
	1000	5.71	0.1436	7.12	0.3321
	average	4.22	0.1073	5.77	0.3107
LCNN+SpecAugment (Park et al., 2019)	1	4.37	0.1094	5.73	0.3082
	100	3.64	0.1080	5.92	0.3218
	1000	4.81	0.1249	6.68	0.3272
	average	4.27	0.1121	6.11	0.3190
LCNN+InF(without consistency loss)	1	2.40	0.0712	3.75	0.2711
	100	2.58	0.0700	4.10	0.2835
	1000	2.95	0.0850	4.27	0.2909
	average	2.59	0.0742	4.04	0.2818

suboptimal on the 2021 dataset. And DFSincNet utilizes Rawboost (Tak et al., 2022) data enhancement, which we do not use.

The SpecAugment method, similar to the InF method, enhances the data by randomly masking in both the time and frequency domains, directly setting certain values to zero. Fig. 3 illustrates the differences between InF and the SpecAugment method:

- InF directly processes the raw waveform through the filter, which can be applied to the input of more models.
- Since InF operates directly on the original waveform, concatenating the first-order dynamic and the second-order features will produce the same masking effect. However, SpecAugment masks the extracted handcrafted features along the time and frequency dimensions, so it cannot produce the same masking effect for the three concatenated features.
- SpecAugment directly assigns 0 to the masked part. However, InF only weakens the other features and highlights the passband features. The augmentation effect of InF is more moderate and does not produce unrealistic data samples.

Table 3 presents the results of the SpecAugment and InF methods. It can be seen from the experimental results that the improvement of InF is stable. SpecAugment does not work well, which we conjecture is related to the size of the masking range.

5. Conclusion

In this paper, we propose a novel model-agnostic training approach which can be seamlessly incorporated into existing SSD models. The InF module utilizes the MSM and KL consistency loss to learn the local context about frequency. And the InI module can learn the global context of multiple instances by cosine similarity loss. Experiments show that both modules can stably improve the detection of the models. In future work, we expect to explore the potential application of our training method in not only SSD but also any task related to speech.

Acknowledgments

This work was supported by the National Nature Science Foundation of China under Grant 62176106, the Special Scientific Research Project of the School of Emergency Management of Jiangsu University under Grant KY-A-01, the Project of Faculty of Agricultural Engineering of Jiangsu University under Grant NGXB20240101, the Key Project of National Nature Science Foundation of China under Grant U1836220, the Jiangsu Key Research and Development Plan Industry Foresight and Key Core Technology under Grant BE2020036.

References

Feng Chen, Shiwen Deng, Tieran Zheng, Yongjun He, and Jiqing Han. Graph-based spectro-temporal dependency modeling for anti-spoofing. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096741. URL https://doi.org/10.1109/ICASSP49357.2023.10096741.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *Int. J. Comput. Vis.*, 132(1):208–223, 2024. doi: 10.1007/S11263-023-01852-4. URL https://doi.org/10.1007/s11263-023-01852-4.

Rohan Kumar Das. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: Asvspoof 2021. *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 29–36, 2021.

Rohan Kumar Das, Jichen Yang, and Haizhou Li. Data augmentation with signal companding for detection of logical access attacks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6349–6353. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9413501. URL https://doi.org/10.1109/ICASSP39728.2021.9413501.

Lijian Gao, Qirong Mao, and Ming Dong. Joint-former: Jointly regularized and locally down-sampled conformer for semi-supervised sound event detection. In *Proc. INTERSPEECH*, pages 2753–2757, 2023.

Lijian Gao, Qirong Mao, and Ming Dong. On local temporal embedding for semi-supervised sound event detection. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:1687–1698, 2024. doi: 10.1109/TASLP.2024.3369529. URL https://doi.org/10.1109/TASLP.2024.3369529.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Bingyuan Huang, Sanshuai Cui, Jiwu Huang, and Xiangui Kang. Discriminative frequency information learning for end-to-end speech anti-spoofing. *IEEE Signal Process. Lett.*, 30: 185–189, 2023. doi: 10.1109/LSP.2023.3251895. URL https://doi.org/10.1109/LSP.2023.3251895.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6367–6371. IEEE, 2022.

Lina J Karam and James H McClellan. A multiple exchange remez algorithm for complex fir filter design in the chebyshev sense. In *Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS'94*, volume 2, pages 517–520. IEEE, 1994.

Changhwan Kim, Se-yun Um, Hyungchan Yoon, and Hong-Goo Kang. Fluenttts: Text-dependent fine-grained style control for multi-style tts. In *INTERSPEECH*, pages 4561–4565, 2022.

Juntae Kim and Sung Min Ban. Phase-aware spoof speech detection based on res2net with phase network. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096672. URL https://doi.org/10.1109/ICASSP49357.2023.10096672.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Tomi Kinnunen, Héctor Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, et al. Tandem assessment of spoofing countermeasures and automatic speaker verification: Fun-

damentals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2195–2210, 2020.

Changtao Li, Feiran Yang, and Jun Yang. The role of long-term dependency in synthetic speech detection. *IEEE Signal Processing Letters*, 29:1142–1146, 2022.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Xiaohui Liu, Meng Liu, Longbiao Wang, Kong Aik Lee, Hanyi Zhang, and Jianwu Dang. Leveraging positional-related local-global dependency for synthetic speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Andreas Nautsch, Xin Wang, Nicholas W. D. Evans, Tomi H. Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md. Sahidullah, Junichi Yamagishi, and Kong Aik Lee. Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(2):252–265, 2021. doi: 10.1109/TBIOM.2021.3059479. URL https://doi.org/10.1109/TBIOM.2021.3059479.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14691–14701. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01430. URL https://doi.org/10.1109/CVPR52688.2022.01430.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. Prompt perturbation consistency learning for robust language models. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 1357–1370. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-eacl.91.

You-cai Qin, Qinghua Ren, Qirong Mao, and Jingjing Chen. Multi-branch feature aggregation based on multiple weighting for speaker verification. *Comput. Speech Lang.*, 77: 101426, 2023. doi: 10.1016/J.CSL.2022.101426. URL https://doi.org/10.1016/j.csl.2022.101426.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas W. D. Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. *Proc. IEEE Int. Conf. Acoust.,Speech Signal Process.*, pages 6369–6373, 2020.

Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Graph attention networks for anti-spoofing. *arXiv preprint arXiv:2104.03654*, 2021.

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE, 2022.

Anton Tomilov, Aleksei Svishchev, Marina Volkova, Artem Chirkovskiy, Alexander Kondratev, and Galina Lavrentyeva. Stc antispoofing systems for the asvspoof2021 challenge. In *Proc. ASVspoof 2021 Workshop*, pages 61–67, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

Jinbo Wen, Jiawen Kang, Minrui Xu, Hongyang Du, Zehui Xiong, Yang Zhang, and Dusit Niyato. Freshness-aware incentive mechanism for mobile ai-generated content (aigc) networks. In *2023 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 1–6. IEEE, 2023.

Jun Xue, Cunhang Fan, Jiangyan Yi, Chenglong Wang, Zhengqi Wen, Dan Zhang, and Zhao Lv. Learning from yourself: A self-distillation method for fake speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

Jiadi Yao, Xing Chen, Xiao-Lei Zhang, Weiqiang Zhang, and Kunde Yang. Symmetric saliency-based adversarial attack to speaker identification. *IEEE Signal Process. Lett.*, 30:1–5, 2023. doi: 10.1109/LSP.2023.3236509. URL https://doi.org/10.1109/LSP.2023.3236509.

Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang. The effect of silence and dual-band fusion in anti-spoofing system. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 4279–4283. ISCA, 2021. doi: 10.21437/INTERSPEECH.2021-1281. URL https://doi.org/10.21437/Interspeech.2021-1281.

Yuxiang Zhang, Zhuo Li, Jingze Lu, Wenchao Wang, and Pengyuan Zhang. Synthetic speech detection based on temporal consistency and distribution of speaker features. *arXiv preprint arXiv:2309.16954*, 2023.